COMPUTAÇÃO EM *GRID* COMO FERRAMENTA ESSENCIAL NA INVESTIGAÇÃO DA ESTRUTURA DO UNIVERSO

ALEXANDRE SUAIDE

MARCELO GAMEIRO MUNHOZ

RESUMO

Movidos pela curiosidade e a necessidade humana de compreender o mundo que nos cerca, a física cria diversas demandas por novas tecnologias. Um exemplo dessa demanda é o processamento de dados gerados em grandes aceleradores de partículas, como o LHC (Large Hadron Collider) recentemente construído no laboratório europeu Cern (European Organization for Nuclear Research). Computação em *grid* consiste em um sistema federativo descentralizado de recursos computacionais, interconectados, com objetivos comuns. A computação em *grid* é um excelente exemplo de produto criado por essa demanda, que certamente terá dezenas, se não milhares, de aplicações em um futuro próximo.

Palavras-chave: física, LHC, Cern, *grid*.

ABSTRACT

Driven by curiosity and by the human need to understand the world around us, physics has generated a broad range of demands for new technologies. One example of such demands is handling the data produced by large particle accelerators, such as the LHC (Large Hadron Collider) *recently built by the European laboratory Cern (European Organization for Nuclear Physics). Grid computing consists of a decentralized federated system of computing resources which are interconnected and share common objectives. Grid computing is an excellent example of a product created by this demand, which will certainly have dozens – maybe thousands – of applications in a near future.*

Keywords: *physics, LHC, Cern, grid.*

Uma das principais características do ser humano, essencial para sua sobrevivência, é a sua capacidade de questionar a origem e a constituição do mundo a sua volta. Essa capacidade permite a geração de conhecimento que resulta no desenvolvimento de métodos e tecnologias que contribuem para o aumento da qualidade de vida da nossa espécie no ambiente em que vivemos. Há muitas maneiras diferentes de procurar pelas respostas para esse questionamento que fazemos diariamente. Do ponto de vista biológico, abordamos essas perguntas sob os aspectos da origem e evolução das espécies e da constituição bioquímica dos organismos presentes na natureza. Do ponto de vista químico, procuramos entender os métodos e mecanismos que permitem a criação, a partir de elementos simples, de moléculas e suas interconexões. Sob o aspecto humano, filosófico, o estudo da evolução do pensamento e da organização social permite entender como progredimos e aprendemos a nos organizar em comunidades estruturadas, coletivas. A física, como ciência natural das mais fundamentais, aborda esses assuntos procurando entender como a matéria que compõe o universo foi formada e como esse universo evoluiu desde sua origem. Procura entender também como essa matéria é constituída na sua forma mais elementar e as relações de interação entre os diversos elementos que a constituem e suas propriedades.

Do ponto de vista da física, toda a matéria conhecida do universo é formada por um pequeno punhado de partículas microscópicas, aparentemente fundamentais. De acordo com o modelo padrão, teoria[1] em física que estuda essas partículas, temos apenas dezoito partículas fundamentais[2]. Ainda mais impressionante é pensar que quase a totalidade da matéria conhecida pode ser resumida à combinação de cinco dessas dezoito partículas fundamentais. Acontece que, mesmo compreendendo tão detalhadamente a estrutura da matéria, ainda há muitas questões em aberto: como o universo surgiu e evoluiu até os dias de hoje? Por que há muito mais matéria que antimatéria no universo? Por que as partículas possuem massa? Só existem três dimensões espaciais? As partículas que conhecemos como fundamentais são, de fato, fundamentais?

Para tentar responder várias dessas perguntas, foi construído, ao longo das últimas duas décadas, no laboratório europeu Cern[3] (European Organization for Nuclear Research), localizado na fronteira entre

ALEXANDRE SUAIDE e **MARCELO GAMEIRO MUNHOZ** são professores do Instituto de Física da Universidade de São Paulo.

1 Em física, uma teoria não tem o caráter especulativo, como em algumas áreas do conhecimento. Uma teoria é uma consolidação sistemática de conceitos e ideias baseada fortemente em observações experimentais de fenômenos naturais.

2 Se considerarmos o bóson de Higgs, a provável partícula recentemente descoberta no LHC, e excluirmos as antipartículas.

3 Ver: https://home.web.cern.ch.

Suíça e França, o acelerador de partículas LHC (Large Hadron Collider). O Cern é um laboratório dirigido por vinte nações europeias com a missão de desenvolver pesquisa em ciências básicas, principalmente física de partículas, aplicações tecnológicas, estimular a colaboração entre cientistas do mundo todo e criar programas educacionais. Com um corpo de funcionários que chega a 2.400 pessoas e mais de 10.000 pesquisadores colaboradores de 113 países, o laboratório agrega em torno de 600 instituições do mundo todo. Com investimentos da ordem de bilhões de euros, o Cern é um excelente exemplo da simbiose entre ciência fundamental e inovação tecnológica e uma demonstração do reconhecimento das nações desenvolvidas sobre a importância dessa relação entre ciência e tecnologia.

Um acelerador de partículas consiste em um equipamento que gera colisões entre diferentes tipos de partículas, sejam elas fundamentais (como os elétrons) ou não (como núcleos de ouro ou chumbo). A partir da medida dos produtos desses choques microscópicos, é possível estudar as propriedades mais fundamentais dessas partículas e a maneira como elas interagem entre si. O LHC é o maior acelerador já construído pela humanidade e envolve um grande complexo de outros aceleradores instalados no Cern, podendo colidir prótons (um dos constituintes básicos do núcleo atômico, ao lado do nêutron) até energias de 14 TeV[4] e núcleos do átomo de chumbo a energias de 1.144 TeV (5,5 TeV por próton ou nêutron que constitui o núcleo desse elemento). Em termos de velocidade, essas partículas viajam a 99,9999991% da velocidade da luz em sentidos opostos em dois anéis circulares de 27 km construídos a, aproximadamente, 150 m da superfície. Em alguns pontos específicos, esses feixes se cruzam, permitindo a colisão entre as partículas que os compõem. Ocorrem cerca de 600 milhões de colisões por segundo. Quando colidem, a energia liberada é suficiente para produzir condições extremas de temperatura e densidade de energia, similares àquelas presentes no universo primordial, apenas alguns microssegundos após o chamado *Big Bang*. A temperatura na região das colisões alcança 10^{12}°C, cerca de 100 mil vezes mais quente que o núcleo do Sol. Nessas condições, podemos explorar quais eram as propriedades físicas nos instantes iniciais da evolução do universo e ter acesso às estruturas mais fundamentais da sua composição.

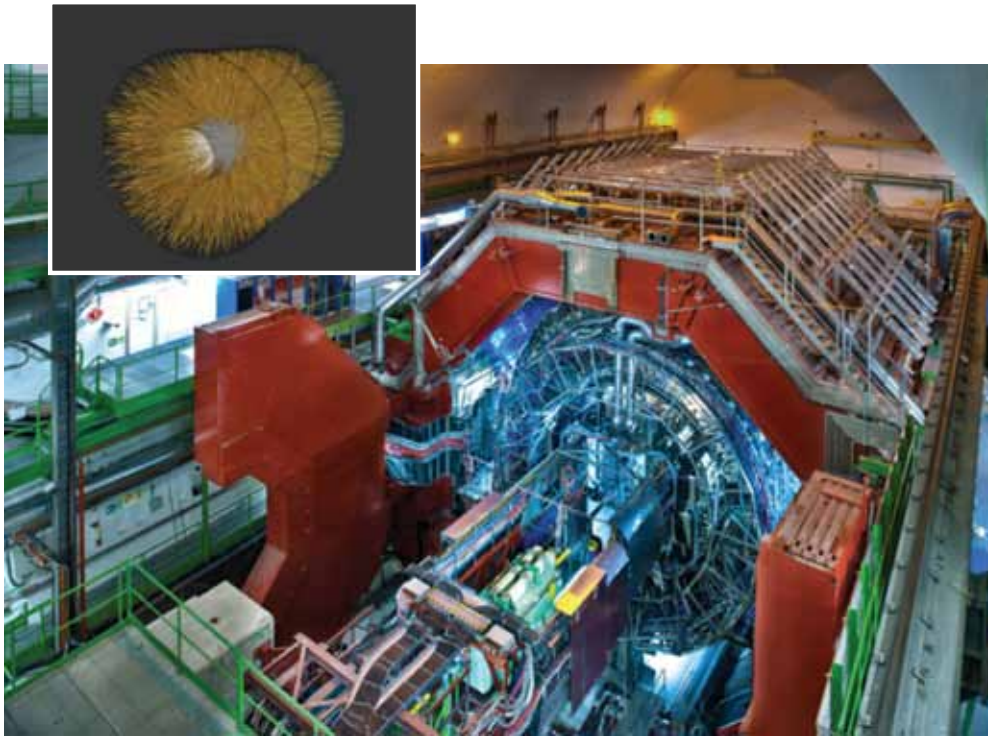
De modo a observar os produtos dessas colisões, nos pontos de cruzamento desses feixes de partículas são montados grandes detectores, que funcionam como gigantescas máquinas fotográficas digitais em 3D que registram a maioria das partículas produzidas e suas propriedades. No total, há quatro grandes experimentos construídos e mantidos por colaborações compostas de milhares de cientistas do mundo inteiro, prontos para analisar os dados gerados por essas "fotografias", que devem ser armazenados e disponibilizados durante muitos anos. O LHC, através de seus experimentos, produz cerca de 1 TB de dados por segundo (cerca de 10 mil Enciclopédias Britânicas por segundo). Em um ano de operação, isso equivale a cerca de 10 PB de dados, ou cerca de 1,5 milhão de CDs de dados anualmente. Como armazenar, processar e analisar esses dados de forma eficiente? Como preservar esses dados por anos, se não décadas, de forma segura? Esse foi, e continua sendo, um grande desafio na área de tecnologia da informação e processamento de dados que motiva dezenas de grupos de pesquisa em grandes universidades mundo afora, incluindo a USP, que participa de dois dos quatro grandes experimentos do LHC: os experimentos Alice e Atlas.

Para permitir o armazenamento, o processamento e a análise desse enorme volume de dados, foi criado o WLCG (Worldwide LHC Computing Grid), um projeto colaborativo global, com mais de 170 centros computacionais em 36 países, que é hoje o maior projeto de computação em *grid* no mundo. Computação em *grid* consiste em um sistema federativo descentralizado de recursos computacionais, interconectados, com objetivos comuns. O que difere sistemas de *grid* de su-

4 A unidade de energia TeV corresponde a, aproximadamente, $1{,}6\times10^{-7}$ J. Apesar de ser uma quantidade pequena de energia do ponto de vista macroscópico, considerando-se que ela está concentrada em uma região do espaço de, aproximadamente, 10^{-15} m de raio, a densidade de energia é bastante alta.

Foto aérea de Genebra mostrando por onde passa o LHC e a localização dos seus experimentos. À direita, pista de pouso e decolagem do Aeroporto Internacional de Genebra. A linha amarela corresponde à fronteira entre França e Suíça.

Fonte: http://aliceinfo.cern.ch



O experimento Alice, com 26 m de comprimento e 16 m de altura e largura, pesa cerca de 10 mil toneladas e funciona como uma gigantesca máquina fotográfica digital em três dimensões, registrando as partículas produzidas quando os feixes colidem. No detalhe, o registro de uma dessas colisões, em que são produzidas milhares de partículas, cada uma representada por uma linha amarela.

percomputadores convencionais, ou *clusters* de processamento, é o fato de um sistema de *grid* ser heterogêneo, geograficamente disperso e fracamente interconectado, ou seja, a comunicação entre os diversos estados que formam a federação é feita através de redes convencionais como a Internet. A descentralização que o *grid* propicia constitui sua grande vantagem, em vários aspectos. Do ponto de vista estratégico, a descentralização faz com que o *grid* não "saia do ar", assim como a Internet. Sempre haverá um *site* disponível com redundância de informação. Do ponto de vista econômico, é a forma mais eficiente de diluir custos, tanto os de aquisição de equipamentos como os de operação (infraestrutura, energia, salários). Do ponto de vista do conhecimento, é uma forma eficiente de distribuir informação tecnológica, uma vez que, por ser operado a partir de centros do mundo inteiro, o conhecimento técnico deve ser rapidamente disseminado, e soluções para problemas podem surgir em vários locais distintos. A possibilidade de ser heterogêneo faz com que, dentro de um sistema de *grid*, um estado possa contribuir com apenas um computador *desktop* enquanto outro estado contribui com um enorme *cluster*, com milhares de CPUs de processamento. A gerência e a comunicação entre os vários estados da federação são feitas através de vários *softwares* que compõem o *middleware* do *grid*. Esse *middleware* gerencia o compartilhamento dos recursos disponíveis entre os usuários e as organizações virtuais. Um estado da federação pode fazer parte de diversas organizações virtuais, alocando recursos computacionais para cada uma delas de acordo com sua política interna.

Outra forma de computação distribuída bastante presente em diversas áreas da física é a chamada computação voluntária. Nesse tipo de processamento de dados, voluntários doam tempo ocioso de processamento de seus equipamentos para ser utilizado em algum projeto científico. Talvez o exemplo mais conhecido de computação voluntária seja o SETI@home[5], acrônimo para Search for Extraterrestrial Intelligence, projeto conduzido pelo Laboratório de Ciências Espaciais da Universidade da Califórnia, em Berkeley, EUA. Esse projeto atingiu milhões de participantes em mais de duas centenas de países. Além de ajudar no processamento de dados de projetos científicos, a computação voluntária corresponde a um excelente meio de divulgação e difusão da ciência. A partir da plataforma desenvolvida para esse projeto, chamada Boinc (Berkeley Open Infrastructure for Network Computing), diversos outros projetos foram desenvolvidos, inclusive no laboratório Cern, onde se criou o LHC@home. Através desse projeto, voluntários geram simulações de processos relacionados ao funcionamento técnico do acelerador LHC e de colisões entre partículas. O que difere projetos desse tipo do *grid* tradicional é a necessidade de maior conhecimento técnico para configurar e operar um estado do *grid*. Isso faz com que seja capacitada uma enorme força de trabalho e pesquisa na área de tecnologia da informação.

O WLCG é estruturalmente organizado de forma hierárquica, em quatro níveis, denominados *tiers*. O *Tier-0* é localizado no centro de computação do Cern. Todos os dados produzidos no LHC passam por esse *tier*, que funciona como um centro de distribuição e é responsável por manter uma cópia de todos os dados brutos adquiridos, pelo primeiro passo de reconstrução dos eventos medidos e pela distribuição de dados para os *Tiers-1*. Apesar do papel fundamental desse *tier* no WLCG, ele é responsável por apenas 20% de todo o poder de processamento do WLCG.

Os *Tiers-1* (11 estados no total) são grandes centros de computação com elevado poder de armazenamento. São responsáveis pelo armazenamento colaborativo dos dados brutos e reconstruídos, dados simulados e pela distribuição dos mesmos para os *Tiers-2*.

Tiers-2 são tipicamente centros de universidades e laboratórios de pesquisa com capacidade de prover espaço de armazenamento e processamento para análise dos dados, de forma proporcional à participação de seus membros nos experimentos

5 Ver: http://setiathome.berkeley.edu.

CENTROS DE PROCESSAMENTO DO EXPERIMENTO ALICE DO LHC NO WLCG

Fonte: alimonitor.cern.ch

do LHC. O WLCG é constituído de aproximadamente 140 desses centros.

Há ainda os *Tiers-3*, que consistem de recursos computacionais oferecidos individualmente por pesquisadores e grupos de pesquisa, sem requerimento mínimo (podendo ser um simples computador pessoal) e comprometimento com disponibilidade.

A Universidade de São Paulo, através do Griper (Grupo de Íons Pesados Relativísticos)[6] contribui desde 2005[7] no desenvolvimento de tecnologias da computação em *grid*, tendo sido um dos grupos pioneiros nessa atividade na área de física nuclear e de partículas no Brasil. Em 2007, por meio de financiamento da Fapesp, o Griper iniciou suas atividades de pesquisa no LHC, com a aquisição de um *cluster* de processamento capaz de prover as necessidades básicas para que a USP participasse do WLCG como um centro do tipo *Tier-2* para processamento de dados do experimento Alice do LHC[8]. A USP vem processando dados para o Alice desde 2009, sendo o único centro no Brasil que oferece suporte a esse experimento. Outros centros no Brasil dão suporte a outros experimentos.

Recentemente, em colaboração com o projeto CloudUSP, um moderno *data center*, com infraestrutura capaz de suportar milhares de CPUs de processamento e cerca de 2 PB de espaço em disco, vem sendo construído no Departamento de Física Nuclear do IF-USP, com o objetivo de suprir a crescente demanda computacional requisitada pelos experimentos do LHC. Esse centro deve ser entregue no primeiro semestre de 2013. Pedidos de financiamento para *upgrades* às agências de fomento, se aprovados, permitirão que a USP se coloque entre os dez maiores centros de processamento de dados do WLCG dedicados ao Alice. Esses *upgrades* permitirão, também, que a USP processe dados de outros experimentos do LHC e de

6 Ver: http://sampa.if.usp.br.

7 Ver: http://www.interactions.org/sgtw/2005/0727/star_saopaulo_more.html.

8 Como ainda não houve assinatura formal do *Memorandum of Understanding* entre Brasil e o Cern para dar suporte computacional à participação da USP no WLCG, somos, formalmente, um centro do tipo *Tier-3*

outras colaborações em *grid*. Em um mundo competitivo e globalizado, o papel das ciências básicas para o desenvolvimento social e tecnológico das nações tem se tornado cada vez mais importante. A simbiose entre disciplinas que buscam aprofundar o entendimento da natureza e do universo *per se*, motivadas pela curiosidade e a necessidade humana de compreender o mundo que nos cerca, estimula intensamente a criação de novas tecnologias. A investigação de aspectos fundamentais da natureza cria desafios que só podem ser superados a partir do desenvolvimento da tecnologia. Essa demanda por inovações tecnológicas criada pela ciência básica é única, não encontrando paralelo em pesquisas e desenvolvimentos de caráter puramente tecnológico, que visam aplicações a curto e médio prazo. A computação em *grid* é um excelente exemplo de produto criado por essa demanda, que certamente terá dezenas, se não milhares, de aplicações em um futuro próximo. Um exemplo atual é o chamado MammoGrid[9], que consiste em uma aplicação da tecnologia *grid* para a elaboração de um banco de dados e o processamento de dezenas de milhares de mamografias provenientes de toda a Europa, permitindo que profissionais de saúde possam desenvolver diagnósticos mais precoces de câncer a partir dessas informações. Sem o desenvolvimento da computação em *grid* em pesquisa básica, essa e outras empreitadas dificilmente se tornariam realidade.

9 Ver: http://knowledge-transfer.web.cern.ch/technology-transfer/external-partners/mammogrid.